Separação cega de sinais de fala em sub-bandas utilizando detectores de voz

Paulo Bulkool Batalheiro, Ronaldo Alencar da Rocha e Diego Barreto Haddad

***Resumo*—Este artigo investiga o emprego de detectores de voz como uma etapa de pré-processamento de uma técnica de separação cega de sinais implementada no domínio do tempo, que emprega estatísticas de segunda ordem para a separação de misturas convolutivas e determinadas. Seu algoritmo foi adaptado para realizar a separação tanto em banda cheia quanto em sub-bandas. A ideia principal visa detectar trechos das misturas que contenham atividade de voz, evitando que o algoritmo de separação seja acionado na ausência de voz, promovendo ganho de desempenho e redução do custo computacional.**

Palavras-Chave*—*Separação cega de fontes, detectores de voz, estruturas em sub-bandas, processamento multitaxas.

Abstract— **This article investigates the use of detectors' voice as a preprocessing step of a blind source separation technique implemented in the time domain, employing second order statistics in the separation of convolutive and determined mixtures. This algorithm is adapted to perform the separation both in fullband and in subbands. The main idea aims at detect portions of the mixtures containing voice activity, avoiding that the separation algorithm is triggered in the absence of voice, promoting performance gain and reduced computational cost.**

Keywords*—*Blind source separation, voice detectors, subbands.

I. INTRODUÇÃO

Na captura de sinais, os sensores existentes registram em geral, sinais corrompidos por diversos fatores, tais como distorções, ruídos, atenuações e interferências. A existência de interferências nos registros de um sensor implica a presença de misturas de diferentes fontes. Cabe ressaltar que tais misturas, como no caso de gravações de músicas, podem ser propositais [1].

Técnicas para separação cega de fontes (BSS, do inglês *Blind Source Separation*) têm sido alvo de grande interesse da comunidade científica na última década, sendo empregadas em um grande número de aplicações tais como: sistemas de áudio [2], reconhecimento de fala [3] e comunicação digital [4], entre outros.

A determinação dos trechos de atividade de fala existentes em um sinal de voz, através de detectores de voz, pode ser útil em aplicações de processamento de sinais de voz, visto que a informação existe apenas em alguns trechos (ou regiões) nos quais existe a fala, a qual pode portanto ser considerada descontínua. Por essa razão, uma vez detectados os trechos de atividade de fala nas misturas, podemos aplicar um algoritmo de separação cega em tais trechos, angariando desta forma uma ligeira melhoria de desempenho e redução de custo computacional, pois o algoritmo de BSS é desativado quando a atividade vocal não é detectada.

Este artigo estrutura-se da seguinte forma: na Seção II, apresentamos as configurações das misturas, a estrutura de separação em sub-bandas que emprega um banco de filtros uniforme modulado por cosseno e o algoritmo de separação em sub-bandas no domínio do tempo empregado. Enfocamos o detector de voz utilizado como etapa de processamento na Seção III. Finalmente as Seções IV e V, respectivamente, contemplam as simulações e as conclusões deste artigo.

II. SEPARAÇÃO CEGA DE FONTES

A. Configurações de Mistura

Dentre as configurações lineares de misturas, as mais desafiadoras são as convolutivas (foco deste trabalho), as quais, num contexto de fontes sonoras, levam em conta a reverberação de um ambiente ecoico. Nestes casos, tipicamente filtros de separação de resposta ao impulso finita (FIR, do inglês *Finite Impulse Response*) com milhares de coeficientes são necessários, tornando muito complexa a tarefa de separação. Para resolver este problema, diversos métodos têm sido propostos na literatura. Enquanto alguns destes realizam a BSS no domínio do tempo [5], outros a efetuam no domínio da frequência [6], costumeiramente almejando melhoria de desempenho ou propondo uma redução de complexidade computacional.

Outra classificação de misturas refere-se à relação entre o número de fontes e o número de misturas (ou sensores). Neste aspecto podemos dividi-las em três casos: sobredeterminadas (quando o número de misturas supera o número de fontes), determinadas (quando o número de misturas é igual ao número de fontes) e subdeterminadas (quando o número de misturas é inferior ao número de fontes). Este artigo considera apenas os casos de misturas determinadas.

B. Sistemas de Mistura e Separação

Seja Q o número de fontes, P o número de sensores, $s_q(n)$ a q-ésima fonte, para $q = 1, ..., Q$, e $x_p(n)$ a p-ésima mistura, para $p = 1, ... , P$, sendo n o índice amostral.

Devido à reverberação presente em um ambiente acústico, pode-se considerar que os sinais das fontes $s_q(n)$ são filtrados por um sistema linear de mistura de múltiplas entradas e múltiplas saídas (MIMO, do inglês *Multiple-Input Multiple-Output*), antes de serem capturados pelos sensores. Considerando somente o caso determinado ($Q=P$), as misturas podem ser descritas por:

Paulo Bulkool Batalheiro e Ronaldo Alencar da Rocha, Programa de Pós-Graduação em Engenharia Eletrônica, Departamento de Engenharia e Eletrônica, UERJ, Rio de Janeiro, RJ, CEP 20559-900, Brasil, e-mails: pbb@uerj.br e ronaldoallencar@gmail.com.

Diego Barreto Haddad, CEFET-RJ, Unidade Descentralizada de Nova Iguaçu, Coordenação de Telecomunicações, Nova Iguaçu, RJ, CEP 26041-271, Brasil, e-mail: diegohaddad@gmail.com

$$x_p(n) = \sum_{q=1}^{P} \sum_{\kappa=0}^{U-1} h_{qp}(\kappa) s_q(n-\kappa), \quad (1)$$

onde $h_{qp}(\kappa)$, $\kappa = 0,..., U-1$, representa a resposta ao impulso de um filtro de comprimento U que modela o caminho acústico (eco) da q-ésima fonte até a p-ésima mistura.

O objetivo das técnicas de BSS é encontrar os coeficientes dos filtros de separação, de modo a obter nas saídas do sistema versões filtradas das fontes originais, tendo acesso apenas aos sinais misturados. As saídas (ou estimativas) do sistema de separação podem ser representadas por:

$$y_q(n) = \sum_{p=1}^{P} \sum_{\kappa=0}^{S-1} w_{pq}(\kappa) x_p(n-\kappa), \quad (2)$$

onde $w_{pq}(\kappa)$, para $\kappa = 0,..., S-1$, representa os coeficientes de um filtro de separação de comprimento S.

C. BSS em sub-bandas

A Figura 1 mostra a k-ésima sub-banda de uma configuração BSS com duas fontes e dois sensores (TITO, do inglês *Two Input Two Output*), considerando um banco de filtros uniforme modulado por cosseno (BFMC) de M canais [7], onde o q-ésimo sinal observado $x_q(n)$ é decomposto pelo filtro de análise $g_k(n)$ e decimado por um fator de decimação (FD_k), sendo os sinais $x^k_q(n)$ resultantes aplicados aos filtros de separação $w^k_{qp}(m)$. Os sinais de saída correspondentes são expandidos pelo mesmo fator FD_k e recombinados pelos filtros de síntese $f_k(n)$ para restaurar os sinais de saída em banda cheia (estimativa das fontes).

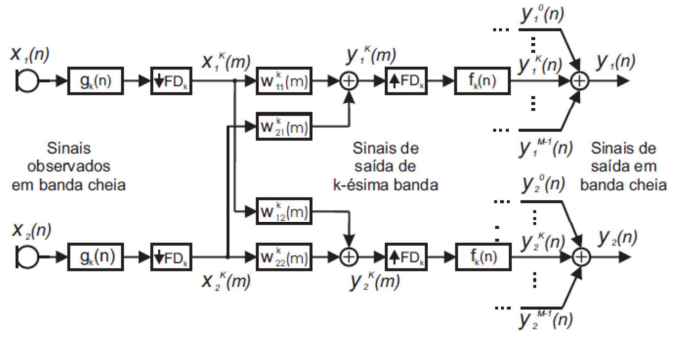

Fig. 1. k-ésimo canal da configuração linear TITO para BSS em sub-bandas.

Supondo um filtro protótipo de comprimento N_p e resposta ao impulso $p(n)$ [8], os filtros de análise e síntese são obtidos da seguinte forma [7]:

$$g_k(n) = 2p(n)\cos\left[\frac{\pi}{M}(k+0{,}5)\left(n-\frac{\Omega}{2}\right)+\theta_k\right],$$
$$f_k(n) = 2p(n)\cos\left[\frac{\pi}{M}(k+0{,}5)\left(n-\frac{\Omega}{2}\right)-\theta_k\right], \quad (3)$$

sendo $\Omega = N_p - 1$ e $\theta_k = (\pi/4)(-1)^k$, para $0 \le k \le M-1$ e $0 \le n \le N_p - 1$.

Para sinais coloridos e não-estacionários, como os sinais de voz, o problema de BSS pode ser equacionado diagonalizando-se diversas matrizes de descorrelação (de tempo curto) das estimativas, considerando múltiplos blocos em diferentes instantes de tempo (TDD, do inglês *Time-Delayed Decorrelation*) [9]. Para efetuar a separação utilizamos o algoritmo *offline* tipo batelada (*batch*) no domínio do tempo, baseado em estatísticas de segunda ordem, proposto em [5], o qual é bastante atrativo pela robustez aos problemas de branqueamento das estimativas e de permutação das saídas, adaptando-o para trabalhar de forma independente em cada sub-banda.

A equação de atualização dos filtros de separação no k-*ésimo* canal é dada por:

$$\mathbf{W}^k(i) = \mathbf{W}^k(i-1) - \frac{2\mu^k(i)}{b^k} \sum_{m=1}^{b^k} \begin{bmatrix} \mathbf{W}^k_{12}\mathbf{R}^k_{21}\left(\mathbf{R}^{-1}_{11}\right)^k & \mathbf{W}^k_{11}\mathbf{R}^k_{12}\left(\mathbf{R}^{-1}_{22}\right)^k \\ \mathbf{W}^k_{22}\mathbf{R}^k_{21}\left(\mathbf{R}^{-1}_{11}\right)^k & \mathbf{W}^k_{21}\mathbf{R}^k_{12}\left(\mathbf{R}^{-1}_{22}\right)^k \end{bmatrix}, \quad (4)$$

onde b_k é o número de blocos utilizados na i-ésima iteração e $\mu^k(i)$ o fator (passo) de aprendizagem, sendo $\mathbf{W}^k_{pq}$ uma matriz do tipo Sylvester, de dimensões $2S_k \times D_k$, definida como:

$$\mathbf{W}^k_{pq} = \begin{bmatrix} w^k_{pq}(0) & 0 & \cdots & 0 \\ w^k_{pq}(1) & w^k_{pq}(0) & \ddots & \vdots \\ \vdots & w^k_{pq}(1) & \ddots & 0 \\ w^k_{pq}(S-1) & \vdots & \ddots & w^k_{pq}(0) \\ 0 & w^k_{pq}(S_k-1) & \ddots & w^k_{pq}(1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & w^k_{pq}(S_k-1) \\ 0 & \cdots & 0 & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \quad (5)$$

$\mathbf{R}^k_{qp}(m) = \left(\mathbf{Y}^k_q(m)\right)^H \mathbf{Y}^k_q(m)$, uma matriz de dimensões $D_k \times D_k$ que contempla, implicitamente, as correlações em diferentes intervalos de tempo (lags) [1], sendo $\mathbf{Y}^k_q(m)$ a matriz $N_k \times D_k$ contendo os D_k blocos atrasados no tempo da q-ésima saída, dada por:

$$\mathbf{Y}^k_q(m) = \begin{bmatrix} y^k_q(mS_k) & \cdots & y^k_q(mS_k - D_k + 1) \\ y^k_q(mS_k + 1) & \ddots & y^k_q(mS_k - D_k + 2) \\ \vdots & \ddots & \vdots \\ y^k_q(mS_k + N_k - 1) & \cdots & y^k_q(mS_k - D_k + N_k) \end{bmatrix}, \quad (6)$$

sendo D_k é o número de blocos atrasados no tempo, na k-ésima sub-banda, contabilizados na estimativa da correlação ($1 \le D_k \le S_k$) e N_k reflete o tamanho de um bloco de sinal de saída ($N_k \ge PD_k$) [5].

Para reduzir a complexidade computacional do algoritmo, o cálculo do fator de normalização $\left(\mathbf{R}^{-1}_{qq}\right)^k$ pode ser simplificado [10], ou seja,

$$\mathbf{R}^k_{qq}(m) \approx \left(\mathbf{y}^k_q(m)\right)^T \mathbf{y}^k_q(m)\mathbf{I}_k \quad (7)$$

sendo sua inversa reduzida ao inverso da potência de um único bloco do sinal de saída (primeira coluna da Eq. (6) e $\mathbf{I}_k$ uma matriz identidade ($D_k \times D_k$).

Um fator de aprendizado adaptativo foi introduzido na Eq. (4) para acelerar a convergência. Os valores do passo de adaptação podem variar (adaptar-se) dependendo da convergência dos coeficientes do sistema de separação a cada iteração. Caso mais do que 70% dos coeficientes dos filtros de separação apresentem em duas iterações consecutivas alterações na mesma direção, na iteração seguinte o valor de μ é escalado por um fator $\alpha=1{,}025$, caso contrário, se o passo for superior ao μ inicial, o fator de aprendizagem, na iteração seguinte, retorna ao seu valor inicial; e por fim caso as duas

condições não sejam verdadeiras, o valor de μ da iteração seguinte é reduzido (dividido) pela mesma constante α.

Em nossas simulações (apresentadas na Seção IV) monitoramos as correlações entre as estimativas das fontes nas diversas sub-bandas com o propósito de evitar problemas de permutação, porém em nenhum caso foi necessário efetuar correções.

D. Complexidade Computacional

Definindo complexidade computacional como o número de multiplicações realizadas a cada iteração (*NMI*) para promover a atualização dos filtros de separação, podemos expressá-la como:

$$NMI = \frac{P^2}{2}\sum_{k=1}^{M} b_k\left(\frac{4S_k^3 + 5S_k^2 + S_k}{FD_k}\right) - 2P\sum_{k=1}^{M} b_k\left(\frac{S_k^2 + S_k}{FD_k}\right), \quad (8)$$

sendo S_k e FD_k, respectivamente, o tamanho dos filtros de separação e o fator de decimação, utilizados no k-ésimo canal, considerando $D_k = L_k$ e $N_k = 2D_k$.

III. Detectores de Voz

Diversos tipos de detectores existentes na literatura foram avaliados, tais como: o Detector Linear e Adaptativo de Energia (ALED, do inglês *Adaptive Linear Energy-Based Detector*) [11], o Detector Linear de Energia em Sub-bandas (LSED, do inglês *Linear Sub-Band Energy Detector*) [11] e o Modelo Estatístico de Detecção de Voz (SMBVAD, do inglês *Statistical Model-Based Voice Activity Detector*) [12]. Conjugando eficiência e baixo custo computacional, utiliza-se neste trabalho o ALED.

A. Detector Linear e Adaptativo de Energia

Este tipo de detector, através de variações na energia do sinal, é capaz de discernir a presença ou a ausência de voz em diferentes trechos do sinal [13], tendo como saída pulsos retangulares enquadrando os trechos com atividade vocal.

Seja *x(i)* a representação da *i*-ésima amostra de um sinal *x(n)*, a implementação do ALED consiste em dividir o sinal em diversos quadros. A energia do *j*-ésimo quadro é dada por:

$$E_j = \frac{1}{K}\sum_{i=(j-1)K+1}^{jK} x^2(i), \quad (9)$$

sendo K o tamanho em amostras do quadro [11].

Em [11] adota-se como critério para a atualização do limiar de detecção:

$$L_{NOVO} = (1-\varphi)L_{ANTERIOR} + \varphi E_{SILÊNCIO}, \quad (10)$$

para $0 < \varphi < 1$, sendo L_{NOVO} o novo limiar a ser calculado, $L_{ANTERIOR}$ o último limiar calculado utilizado, $E_{SILÊNCIO}$ a última energia calculada para o último quadro em que não houve detecção de voz e φ um parâmetro escolhido conforme explicado em [11], cuja importância é explicada em [13].

Para aumentar a robustez, utiliza-se informação estatística de segunda ordem, do conjunto das energias calculadas para os m quadros mais recentes onde não houve detecção de voz, para a atualização de φ. Seja M_n uma memória de m componentes composta pelos valores das energias dos m últimos quadros com ausência de voz. A atualização da memória é realizada no instante em que se detecta ausência de voz em um novo quadro, descartando o valor de energia mais antigo. Em [11] é mostrado como definir o limiar inicial, ou seja, como escolher os m valores iniciais que preenchem a memória M_n.

IV. Resultados Experimentais

Simulações computacionais comparam o desempenho do algoritmo de BSS *offline* implementado no domínio do tempo e adaptado para realizar a separação utilizando ou não detectores de voz, tanto em banda cheia (*M=1*) quanto em sub-bandas. Todos os experimentos foram realizados usando sinais de fala com 10 segundos de duração e frequência de amostragem de 16 kHz.

Para a simulação das misturas foi utilizada uma sala acústica virtual [14] de dimensões 4m × 3m × 2,5m, considerando o caso determinado com duas fontes e dois microfones/sensores (*P=Q=2*), espaçados de 5cm. As fontes sonoras foram posicionadas a 1m de distância do ponto médio entre os sensores, em duas direções diferentes: – 45° e 45°.

Na simulação do sistema de separação cega, foram utilizadas 500 iterações, tempos de reverberação diferentes (T_{60}=25ms e T_{60}=100ms), comprimento dos filtros FIR do sistema de separação $S_k = 256/FD_k$, número de intervalos de tempo empregados nas estimativas de correlação $D_k=S_k$, tamanho dos blocos das saídas utilizados para estas estimativas $N_k=2D_k$. Os filtros de separação $w_{pq}^k(n)$ que foram inicializados conforme [5], ou seja, primeiro coeficiente unitário para $p=q$ e demais zerados para todos os filtros.

O valor inicial dos fatores de aprendizado em todos os casos foi $\mu^k(0)=0.001$ para todas as sub-bandas

No ALED utilizamos memória *m=64*, fator de margem de erro *k = 4*, janela de tamanho *K=32* e os parâmetros apresentados na Tabela I, que foram modificados de [11], por tentativa e erro, para estabilizar melhor a detecção de fala, onde ξ é um parâmetro especificado em [11].

TABELA I. Critério de atualização de φ.

ξ	φ
$\xi \geq 1{,}25$	0,30
$1{,}10 \leq \xi \leq 1{,}25$	0,25
$1{,}00 \leq \xi \leq 1{,}10$	0,15
$\xi \leq 1{,}00$	0,10

Além disso, prolongamos os pulsos retangulares (a saída) do detector de forma a evitar erros proporcionados pela demora na atualização do limiar, evitando desconsiderar blocos que contenham atividade de voz, pois é preferível que o detector reconheça a existência de voz em alguns blocos de silêncio a desconsiderar blocos que possam conter atividade vocal e comprometer o desempenho do algoritmo de separação.

Para as implementações em sub-bandas foram utilizados bancos de filtros modulados por cosseno com *M=2* e *4* bandas, implementados através de filtros protótipos PR com N_p=16M. Os filtros de separação são adaptados de forma independente em cada sub-banda, possibilitando trabalhar com FD_k diferentes em bandas distintas, reduzindo o *aliasing* causado pela atenuação finita dos filtros de análise e síntese e almejando um compromisso entre desempenho e custo computacional. Com esse objetivo, simulamos versões mescladas para *M=4* (*M4*), sendo a primeira banda (B_0) e a quarta banda (B_3) implementadas com $FD_{0\,e\,3}=2$, e a segunda e terceira banda, respectivamente as bandas, B_1 e B_2, realizadas com $FD_{1\,e\,2}=1$; e para *M=2* (*M2*), onde a primeira banda (B_0) é implementada com $FD_0=1$ e a segunda banda (B_1) com $FD_1=2$.

Na avaliação do desempenho das diversas configurações apresentadas utilizam-se as métricas: Relação Fonte-Interferência (SIR, do inglês *Source-to-Interference Ratio*), Relação Fonte-Artefato (SAR, do inglês *Source-to-Artifact-Ratio*) e Relação Fonte-Distorção (SDR, do inglês *Source-to-Distortion-Ratio*), calculadas como descrito em [15], considerando em todos os casos os valores médios entre as métricas das duas estimativas.

As Figuras 2 e 3 apresentam a evolução da SIR média para M=1, 2 e 4, $FD_k=\lceil M/2 \rceil$, tempo de reverberação de 25 ms, misturas sem trechos de silêncio e com trechos de silêncio, sem uso do detector (SD) e com uso detector (CD).

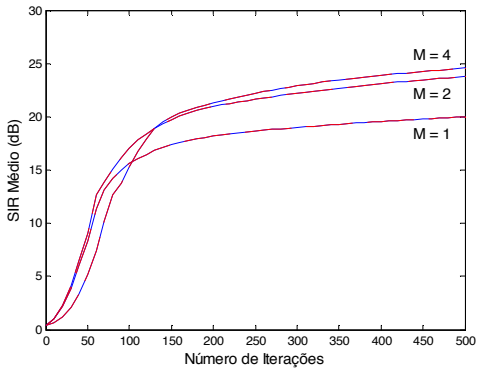

Fig. 2. Evolução da *SIR* média para *M=1, 2* e *4*, $FD_k=\lceil M/2 \rceil$, tempo de reverberação de 25 ms e misturas sem trechos de silêncio, sendo SD em linha tracejada azul e CD em vermelho.

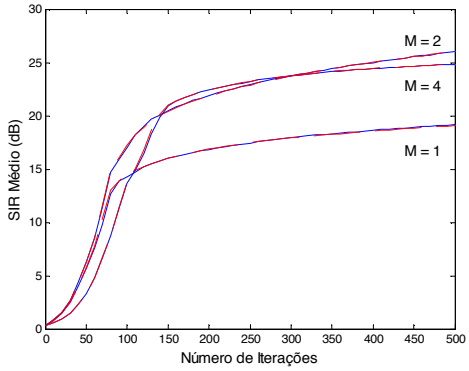

Fig. 3. Evolução da *SIR* média para *M=1, 2* e *4*, $FD_k=\lceil M/2 \rceil$, tempo de reverberação de 25 ms e misturas com trechos de silêncio, sendo SD em linha tracejada azul e CD em vermelho.

TABELA II. RESULTADOS MÉDIOS FINAIS DA SIR, SAR E SDR, PARA DIFERENTES VALORES DE FD_k, T_{60}=25 ms E MISTURAS SEM TRECHOS DE SILÊNCIO.

M	FD_k	SIR		SAR		SDR	
		SD	CD	SD	CD	SD	CD
1	1	20,00	20,00	74,60	74,60	20,00	20,00
2	1	23,80	23,80	70,72	70,72	23,80	23,80
2	2	21,46	21,46	34,96	34,96	20,01	20,01
2	*M2*	22,12	22,12	24,85	24,85	19,33	19,33
4	1	28,33	28,33	73,87	73,87	28,33	28,33
4	2	24,59	24,59	31,38	31,38	23,70	23,70
4	*M4*	26,73	26,73	72,78	72,78	26,73	26,73

As Tabelas II e III mostram os resultados médios finais da SIR, SAR e SDR, para diferentes valores de FD_k e T_{60}=25 ms, e as Tabelas IV e V apresentam os resultados das métricas para diferentes valores de FD_k e T_{60}=100 ms.

A evolução da SIR média para *M=1, 2* e *4*, $FD_k=\lceil M/2 \rceil$, com tempo de reverberação de 100 ms, para misturas sem trechos de silêncio e com trechos de silêncio, sem uso do detector (SD) e com uso detector (CD), por questões de similaridade com as figuras 2 e 3 foram omitidas neste artigo.

TABELA III. RESULTADOS MÉDIOS FINAIS DA SIR, SAR E SDR, PARA DIFERENTES VALORES DE FD_k, T_{60}=25 ms E MISTURAS COM TRECHOS DE SILÊNCIO.

M	FD_k	SIR		SAR		SDR	
		SD	CD	SD	CD	SD	CD
1	1	19,16	19,06	45,41	45,42	19,14	19,05
2	1	25,98	25,99	51,24	50,89	25,87	25,87
2	2	22,22	22,23	31,62	31,62	21,69	21,69
2	*M2*	23,18	23,18	24,95	24,95	20,69	20,68
4	1	29,26	29,27	53,40	53,41	29,05	29,05
4	2	24,81	24,83	33,60	33,53	24,10	24,11
4	*M4*	28,57	28,59	53,28	53,35	28,39	28,40

TABELA IV. RESULTADOS MÉDIOS FINAIS DA SIR, SAR E SDR, PARA DIFERENTES VALORES DE FD_k, T_{60}=100 ms E MISTURAS SEM TRECHOS DE SILÊNCIO.

M	FD_k	SIR		SAR		SDR	
		SD	CD	SD	CD	SD	CD
1	1	14,65	14,65	24,67	24,67	14,22	14,22
2	1	15,21	15,21	24,60	24,60	14,72	14,72
2	2	14,47	14,47	24,16	24,16	13,96	13,96
2	*M2*	14,69	14,69	20,50	20,50	13,54	13,54
4	1	15,48	15,48	24,46	24,46	14,95	14,95
4	2	14,78	14,78	24,10	24,10	14,27	14,27
4	*M4*	15,40	15,40	24,48	24,48	14,88	14,88

TABELA V. RESULTADOS MÉDIOS FINAIS DA SIR, SAR E SDR, PARA DIFERENTES VALORES DE FD_k, T_{60}=100 ms E MISTURAS COM TRECHOS DE SILÊNCIO.

M	FD_k	SIR		SAR		SDR	
		SD	CD	SD	CD	SD	CD
1	1	15,58	15,56	25,90	25,81	15,12	15,09
2	1	16,41	16,42	25,66	25,69	15,85	15,84
2	2	15,67	15,67	24,82	24,82	15,12	15,12
2	*M2*	16,19	16,19	22,07	22,07	15,12	15,12
4	1	17,16	17,21	25,71	25,73	16,53	16,57
4	2	16,44	16,51	25,20	25,09	15,84	15,89
4	*M4*	17,10	17,14	25,68	25,70	16,47	16,51

Os resultados apresentados nas Figuras 2 e 3 (bem como os resultados omitidos para T_{60}=100ms) indicam que estruturas em sub-bandas apresentam desempenho significantemente superior à estrutura em banda cheia. Aumentando os FD_k na implementação em sub-bandas, a BSS é realizada em taxas de amostragem menores que a taxa em banda cheia, reduzindo a complexidade computacional envolvida (ver Eq. (8)), porém aumentando o *aliasing* entre os canais. Observando as Tabelas II, III, IV e V, verifica-se que a elevação dos FD_k promove queda nas métricas das estruturas em sub-bandas em relação aos casos sem decimação (FD_k=1), mas mantendo ainda um desempenho superior aos resultados em banda cheia. Uma opção interessante é usar diferentes FD_k nas diferentes sub-bandas, de modo a prover um bom compromisso entre complexidade e desempenho.

As Tabelas VI e VII indicam o número de blocos utilizados pelo algoritmo de separação e o *NMI*, a cada iteração, para as simulações apresentadas. A análise destes resultados indica que

a etapa de detecção de voz, no caso de BSS em sub-bandas, promove importante redução do número de blocos utilizado pelo algoritmo de separação a cada iteração, principalmente nas bandas de frequência mais altas, na presença de trechos de silêncio nas misturas, acarretando redução do custo computacional envolvido.

TABELA VI. NÚMERO DE BLOCOS E NÚMERO DE MULTIPLICAÇÕES DO ALGORITMO DE SEPARAÇÃO COM $FD_K=\lceil M/2 \rceil$ PARA UM MISTURA SEM TRECHOS DE SILÊNCIO.

M = 1							
T_{60}	Sem Detector		Com Detector				
	B_0	NMI	B_0				*NMI*
25 ms	621	$8{,}36x10^{10}$	621				$8{,}36x10^{10}$
100ms	621	$8{,}36x10^{10}$	621				$8{,}36x10^{10}$
M = 2							
T_{60}	Sem Detector		Com Detector				
	$B_{0,1}$	*NMI*	B_0		B_1		*NMI*
25 ms	621	$1{,}67x10^{11}$	621		621		$1{,}67x10^{11}$
100ms	621	$1{,}67x10^{11}$	621		621		$1{,}67x10^{11}$
M = 4							
T_{60}	Sem Detector		Com Detector				
	$B_{0,1,2,3}$	*NMI*	B_0	B_1	B_2	B_3	*NMI*
25 ms	621	$2{,}10x10^{10}$	621	621	621	618	$2{,}09x10^{10}$
100ms	621	$2{,}10x10^{10}$	621	621	621	618	$2{,}09x10^{10}$

TABELA VII. NÚMERO DE BLOCOS E NÚMERO DE MULTIPLICAÇÕES DO ALGORITMO DE SEPARAÇÃO COM $FD_K=\lceil M/2 \rceil$ PARA UM MISTURA COM TRECHOS DE SILÊNCIO.

M = 1							
T_{60}	Sem Detector						
	B_0	*NMI*	B_0				*NMI*
25 ms	621	$8{,}36x10^{10}$	573				$7{,}71x10^{10}$
100ms	621	$8{,}36x10^{10}$	577				$7{,}77x10^{10}$
M = 2							
T_{60}	Sem Detector		Com Detector				
	$B_{0,1}$	*NMI*	B_0		B_1		*NMI*
25 ms	621	$1{,}67x10^{11}$	616		501		$1{,}50x10^{11}$
100ms	621	$1{,}67x10^{11}$	621		508		$1{,}52x10^{11}$
M = 4							
T_{60}	Sem Detector		Com Detector				
	$B_{0,1,2,3}$	*NMI*	B_0	B_1	B_2	B_3	*NMI*
25 ms	621	$2{,}10x10^{10}$	621	596	490	461	$1{,}83x10^{10}$
100ms	621	$2{,}10x10^{10}$	621	596	492	469	$1{,}84x10^{10}$

Os resultados evidenciam que o emprego de detectores de voz não compromete o desempenho do algoritmo de BSS, mesmo em casos de misturas sem trechos de silêncio, revelando-se importante, quando lidamos com misturas que apresentam instantes de silêncio, não somente pelo desempenho ligeiramente superior, mas principalmente pela redução da complexidade computacional quando empregamos estruturas em sub-bandas.

V. CONCLUSÕES

Neste artigo, a detecção de voz como uma etapa de pré-processamento na tarefa de separação cega de fontes no domínio do tempo foi investigada. Foram utilizadas misturas convolutivas com presença e com ausência de trechos de silêncio, para diferentes condições de reverberação. Para verificação da qualidade das estimativas foram usadas a SIR, SAR e SDR. Simulações computacionais foram realizadas envolvendo sinais de fala, mostrando o desempenho superior das estruturas em sub-bandas em relação à banda cheia, considerando diferentes fatores de decimação. A utilização dos detectores de voz mostrou-se bastante interessante porque o algoritmo de separação teve um desempenho similar nos casos de misturas sem trechos de silêncio e ligeiramente superior quando as misturas continham trechos de silêncio; e para os casos de misturas com ausência de fala promoveu significante redução do número de blocos utilizado a cada iteração e consequentemente redução da complexidade computacional envolvida, principalmente nas bandas mais altas das implementações em sub-bandas, onde a concentração de energia do sinal de voz é menor.

REFERÊNCIAS

[1] M. R. Petraglia, P. B. Batalheiro, and D. B. Haddad, "Métodos de Separação Cega de Fontes", *Tutoriais do XVII Congresso Brasileiro de Automática*, vol. X, n. X, pp. 25, Apr. 2008.

[2] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation", *IEEE Transactions on Speech And Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[3] D. B. Haddad, M. R. Petraglia, and P. B. Batalheiro, Técnicas para Separação Cega de Fontes aplicadas na Melhoria de Desempenho de um Classificador de Palavras Isoladas. *In: 5º. Congresso de Engenharia de Áudio da AES-Brasil, 2007, São Paulo. Anais do 5º. Congresso de Engenharia de Áudio, 2007*. vol. 1. pp. 20-27.

[4] A. T. Erdogan, "Globally convergent deflationary instantaneous blind source separation algorithm for digital communication signals," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 2182–2192, 2007.

[5] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics", *IEEE Trans. on Speech and Audio Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[6] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 70–78, 2007.

[7] P. P. Vaidyanathan, Multirate Systems and Filter Banks. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[8] T. Q. Nguyen, "Digital filter banks design - quadratic constrained formulation", *IEEE Trans. on Signal Processing*, vol. 43, pp. 2103–2108, Sep. 1995.

[9] J. Bourgeois and W. Minker, Time-Domain Beamforming and Blind Source Separation. Spring Street, NY: Springer, 2007.

[10] Buchner, H., Aichner, R. and Kellerman, W. "Blind Source Separation for Convolutive Mixtures: A Unified Treatment", *In Y.Huang and J. Benesty (eds.), Audio SignalProcessing, Kluwer Academic Publishers, Boston*, pp. 40, Feb. 2004.

[11] R. V. Prasad, A. Sangwan, H. S. Jamadagni, M. C. Chiranth, R. Sah, and V. Gaurav, "Comparison of Voice Activity Detection Algorithms for VoIP". *In Proc. 2002 Seventh Int. Symp. On Computers and Communications*, pp. 530 – 535, Jul. 2002.

[12] D. Y. Cho, and A. Kondoz, "Analysis and Improvement of a Statistical Model-Based Voice Activity Detector". *IEEE Signal Processing Letters*, vol. 8, no 10, pp. 276-278, Oct. 2001.

[13] F. S. P. Clark, M. R. Petraglia, and D. B. Haddad, "Cancelamento de Eco Acústico e Separação Cega de Fontes Aplicados à Telefonia Viva-Voz", pp. 92, Dec. 2010.

[14] E. A. Lehman and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of The Acoustic Society of America*, vol. 124, no. 1, pp. 269–277, Jun. 2008.

[15] E. Vincent, R. Gribonval, and C. Fvotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.